Project Report

- **Project Objective-** The purpose of this project is to predict whether a customer is going to default or not based on historical data.

- **Dataset Source-** **Default of Credit Card Clients Dataset**

- **Dataset Introduction-** This dataset contains information on default payments, demographic factors, education, sex, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. There are 24 independent features and 1 ‘default.payment.next.month’ is a dependent/ target feature in a dataset. Dataset contains 30000 rows in a data

- **Data Cleaning Techniques-**
  Detected 12% outliers on the basis of some columns , but decided not to remove that because of domain knowledge.
  Null values identification technique used , dataset is absence of null values.
  Replaced values in a column with significant numbers and removed insignificant values.
  Scaling of data to get scaled values.

- **Data Analysis-**
  Analysis of data by exploring it.
  Checked for numerical and character values,6 point summary
  Univariate analysis of discrete data and numerical data

- **Feature Engineering-**Created new columns as follows:
  - BILL_AMT_AVG_Last6m- Average of Bill Amounts of last 6 months
  - PAY_AMT_AVG_Last6m- Average of previous payments of last 6 months
  - Max_Pay_Last6m- Maximum value of previous payments
  - Max_Bill_Amt_Last6m- Maximum value of Bill Amounts
  - SEX_MARRIAGE- Combination of SEX and MARRIAGE columns
  - Credit Utilization Ratio- Ratio of BILL AMOUNT and LIMIT BALANCE

- **Classification Models building-** Applied following models on given dataset :
  - Logistic Regression
  - Decision Tree Classifier
  - Random Forest Classifier
  - Support Vector Classifier
  - Gradient Boosting Classifier
  - XGBoost [eXtreme Gradient Boosting] Classifier

- **Steps taken to improve accuracy of models-** Above mentioned models building on different sets of data:
  - Upsampled dataset
  - Downsampled dataset
  - Dataset built by SMOTE technique

- **Conclusion-**
  Achieved 80% accuracy by Gradient Boosting Classifier applied on dataset built by SMOTE technique.